数 字 与 人 文 研 究

古代文学研究的数据来源、指标和意义

王 兆 鹏

(中南民族大学 文学与新闻传播学院, 武汉 430074)

摘　要: 古代文学研究需要数据化,文献目录和作品文本都是数据来源。书名目录、作品篇名目录、文学史章节目录、文学研究论著目录都可转化为文学数据;作品文本中的文学体裁、主题、人物、评点、研究、传播、接受等也可提取为数据。不同的数据有不同的意义,断代总集的作品目录数据,可以分析文体的发展演进;诗文选本的目录数据,可以用来考察作家作品的影响力;文学史的章节目录数据,反映出作家地位的高低、文学史编写者的文学观念和价值标准及其变化;论著目录数据,可统计分析学术史的发展进程;作家的生活年代和作品数据,可以观察作家的时空分布;作品中的人物数据,可探寻作家的社会关系及其影响力。文学研究数据,也有局限性。定性分析与定量分析的有机结合,才能最大限度地解决文学研究的各种问题。

关键词: 中国古代文学研究;数字人文;量化分析;数据化

中图分类号: G 250.7　**文献标识码:** A　**文章编号:** 1000-5919(2024)06-0078-10

古代文学研究的数据化,是时代的需求、学科发展的需求,也是学术创新的需求。然则,文学数据从哪来,需要哪些数据？数据又有什么用？本文将从文学研究的数据来源、数据指标和数据意义三个方面来讨论。

一、数据来源

文学研究的数据来源有二:一是文献目录,二是作品文本。

先说文献目录。与文学研究相关的目录有四种类型:古籍书名目录、作品篇名目录、文学史章节目录、文学研究论著目录。这四种目录都可转化为数据,甚至可以说,凡目录,皆数据。

文学古籍书名目录,包含在古籍书目中。古籍书目,资源丰富,类型多样,古代有公私藏书目,现代有馆藏书目和专题书目。

古代公藏书目,既见历代正史的《艺文志》和《经籍志》,又有后人增订补编的各代史书艺文志,如姚振宗《汉书艺文志拾补》、倪灿《宋史艺文志补》等。私家藏书目,自南宋尤袤《遂初堂书目》、晁公武《郡斋读书志》和陈振孙《直斋书录解题》以下,更是汗牛充栋。宋明以来私家藏书目,近年有不少丛刊予以汇编,如《宋元明清书目题跋丛刊》《明人书目题跋丛刊》《清人书目题跋丛刊》等。这些丛刊里的书目,包含了海量的书名目录信息,经过技术转化后即可成为数据。

历代公私藏书目,反映的是历史上曾经传藏的书目。要了解当今存世的古籍书目并转化为数据,可以利用三类书目,一类是丛书目录,如《中国丛书综录》等;另一类是单行本书目,如《中国古籍善本书目》《中国古籍总目》《日藏汉籍善本书录》等;又一类是馆藏书目,如国内的《北京大学图书馆藏古籍善本书目》和《台湾公藏善本书目书名索引》《香港所藏古籍书目》,美国的《美国国会图书馆藏中国善本书录》,日本的《内阁文库汉籍分类目录》等。

收稿日期: 2024-04-30

作者简介: 王兆鹏,中南民族大学文学与新闻传播学院教授,四川大学文科讲席教授。

基金项目: 国家社会科学基金重大项目“汉魏六朝文学编年地图平台建设”(19ZDA253)。

文学古籍书目,还可查分类专题书目。断代别集书目,有胡旭《先唐别集叙录》、赵荣蔚《唐五代别集叙录》、祝尚书《宋人别集叙录》、崔建英《明别集版本志》、汤志波和李嘉颖《明别集整理总目》、李灵年和杨忠《清人别集总目》等。分体书目,有饶宗颐《词籍考》,吴熊和、严迪昌和林玫仪合编《清词别集知见目录汇编》,石昌渝《中国古代小说总目》,庄一拂《古典戏曲存目汇考》等。每类书目,著录的宗旨和原则不同,收录的范围对象也不同,需要相互补充、交相为用,仅靠某类书目很难将需要的书目数据搜罗完备。

作品篇名目录数据,包含在各种总集和别集中。总集,有三大类:一类是断代总集,如《先秦汉魏南北朝诗》《全唐诗》《全唐文》《全唐五代词》《全宋词》《全宋诗》《全宋文》《全金元词》《全元文》《全元诗》《全元散曲》《全元戏曲》《全明散曲》《全明戏曲》《全清散曲》等;另一类是地方性总集,如《全闽词》《全蜀词》《全粤诗》《全台诗》《全台词》等;又一类是总集型选本,如《文苑英华》《唐文粹》《乐府雅词》《唐诗别裁》《宋词三百首》之类。别集,也有两类:一类是全集型,如《陶渊明集》《杜甫全集校注》《苏轼文集编年笺注》等;一类是选集型,如《杜甫诗选》《李白诗选》《苏轼选集》等。书名目录和作品目录,都可以转化为有用的数据。

文学史章节目录,也是数据来源。因为文学史章节目录,基本上是以作家为中心,不同作家在同一文学史里所占章节份额不同,同一作家在不同文学中所占章节份额不同,不同的章节份额构成了作家地位的差异性。20 世纪出版了数千种中国文学史著作。陈飞《中国文学专史书目提要》(大象出版社 2004 年版)就搜罗了 2000 年以前出版的 2885 部中国文学史书目。其中正目收录 710 部文学史的章节目录,极便于利用。如按图索骥,将二千多种中国文学史著作章节目录进行转化,可以获得大量作家文学史地位的数据。

研究论著目录,一般是把它作为查询文献的线索,其实它也有学术史价值。汇集一个学科或一个领域的研究论著目录,能够反映学术史的发展演进。方建新编《二十世纪宋史研究论著目录》,就是为了“反映 20 世纪宋史研究的学术发展史。即既要反映上个世纪宋史研究总的情况,又要通过这一目录,揭示上个世纪宋史研究的各个方面、各个历史阶段、不同地区(大陆、港台)宋史研究发展的概况、轨迹、特点、热点。”①因此,研究论著目录,完全可以转化为计量学术史的数据,以统计分析一个时期、一个领域研究热点的分布和时段变化、作者队伍的构成和更替等,全面考察学术史的发展演进。

文学研究论著目录的来源,有两个方面,一是利用学界整理的论著目录成果进行转化,如台北五南图书出版公司 1996 年出版的《中国文学论著集目》系列,李南晖主编《中国古代文体学论著集目(1900—2014)》(北京大学出版社 2016 年版),钱振民主编、曹鑫编著《20 世纪中国古代文学研究文献总目 · 明代论文卷》(国家图书馆出版社 2021 年版)等。二是自行搜集整理。可根据《全国总书目》和全国报刊资料索引(https://www.cnbksy.com/)、中国人民大学书报资料中心官网(http://zlzx.ruc.edu.cn/)、中国知网(https://www.cnki.net/)和中国国家图书馆官网(https://www.nlc.cn/)等网站搜辑有关研究论文和著作目录。

需要注意的是,无论是已有的论著目录,还是通过网站查询,任何一种论著目录索引或网站都不可能将自己需要的目录一网打尽,必须辑录汇聚多种目录信息来源,相互补充,才可能臻于完备。

已有的论著目录索引,分类目录最便于转化和统计分析。因为分类目录,是编者根据学科或领域的知识体系进行分类编排,每个类属就是一个专题或一个研究方向,便于数据的归类统计分析。标引时也很便捷,按类标引即可,不必自己费力区分。如曹鑫编著的《20 世纪中国古代文学研究文献总目 · 明代论文卷》除作家作品外,又有总论,总论下分综论、文学史、文论、诗词曲、文赋、小说戏曲、民间文学、少数民族文学、文学与宗教、比较文学、文学流派等类。转化为数据后,便于对这些专题研究进行分类统计

① 方建新:《二十世纪宋史研究论著目录》,北京:国家图书馆出版社 2006 年版,第 1901 页。

分析,考察每个研究专题的发展过程。

再说作品文本。作品,既包括历代的文学作品,也包含后世评点、研究文学的论著。作品文本的数据转化,有三种方式:一是将作家小传信息转化为作家基本数据,二是将作家生平事迹转化为作家活动和创作编年数据,三是从作品内容中挖掘提取数据。

经过整理校点的作品集,都有作家小传或生平基本信息的介绍。无论是《全唐诗》《全宋词》《全元文》之类的断代总集,还是《唐诗选》《宋词选》之类的作品选集,抑或是《李白诗选》《杜甫诗选》之类的别集型选本,甚至是文学家辞典如《中国文学家大辞典》之类的工具书,都包含有作家生平的基本信息,可以从中挖掘提取作家生平的基本数据。

作家活动和创作编年的数据来源,有两类文献:一是研究论著,如作家年谱和别集编年笺注、作家生平和作品编年考订的论文等,二是传记史料,如正史中的传记,编年史中的作家仕履事迹,文集中的行状、神道碑、墓志铭等个人传记和书简序跋等。有今人所著年谱和考订的论著,就尽量以之为依据来挖掘作家生平活动和作品编年数据。如果有些作家没有相关的生平和作品研究论著,就只好通过古代的传记资料来挖掘和提取。

二、数据指标

了解了古代文学研究的数据来源,再来讨论其数据指标,即文学研究需要哪些类型、性质的数据。不同文献和不同研究目的指向,需要的数据指标不尽相同。

书名目录数据,应包含书名、作家、卷数、版本等数据指标。作家的生平信息数据,一般书目题跋记载都比较简略,可以融合其他数据库来补全所需的作家数据指标。版本数据指标,又可细分为编辑整理者、刊刻者、校订者、题跋者、收藏者、成书时间、刊刻年代、刊刻地点等。书目数据需要注明数据来源,包含书目的作者、书名、版本和页码,以便校核。

作品篇名目录数据指标,应包括作品篇名、作者、所属文体、创作的时间和地点、创作的对象或相关人物。除作品篇名外,其他数据需要标引。作品所属文体,有的文章自带文体标识,如表、奏、记、序、墓志、碑铭之类,有的则需要人工辨识标引;赋体从篇名上可以辨别,但属大赋还是小赋,是骈赋还是文赋,也需要人工标引。诗体的层次还需细分(详下)。如何分类,取决于量化分析时的需求,可粗分也可细分,先分大类再分小类。古籍文献知识图谱台(https://www.cnkgraph.com/)对所录诗歌的体式已作区分标识,可以参考取用。至于词作,调名之下有题序者需引录题序,无题序者引首句,以为区分。诗歌特别是乐府诗,也有同题作品,如《出塞》《从军行》《行路难》之类,也需引录首句以作区别。篇名数据,同样应明确标注其来源,如书名、作者、版本和页码等。

文学史章节目录的数据指标,应包含作家、作品两大类。作家,可分专章、合章、专节、合节、附及五等。以袁行霈先生主编《中国文学史》唐代部分为例,李白、杜甫、李商隐三人各为专章;韩孟诗派与刘禹锡、柳宗元,白居易与元白诗派等分别合为一章;陈子昂、李贺各占一节,是为专节;王绩与四杰、杜审言与沈宋、王维和孟浩然、高适和岑参、顾况和李益等分别合为一节,是为合节。所谓附及,是指有些作家仅在正文里附带述及,其名字不在章节目录中出现,所占篇幅很少,如吕温、李德裕,是在论述刘禹锡、柳宗元时提及,各占一小段的篇幅;李璟,则是在论述冯延巳和李煜时附带提及,评述的篇幅仅百余字。作家在文学史章节中等次的不同,意味着其文学史地位高低的差异。文学史中论述的作品,有两种情形,一类是立专节或合节论述的作品集,如《水经注》《洛阳伽蓝记》《世说新语》《花间集》等,这类作品集,如同重要作家,出现在章节目录中,可参照作家的同等情况区分其等次。另一类是论述作家创作特色时例举的代表作,这类代表作又可分三个等次:第一等是全引作品原文、或系长篇而节引部分段落并有评述;第二等是列举篇名并节引部分文句或诗句;第三等仅提及篇名。引述作品的情况不同,显示出作品的重要性不同。文学史在论述作家的内容特色、艺术贡献时,列举哪些作品、如何分析评价作品,都

是经过慎重考量,既显示出文学史家对作品的认识和评价,也折射出他们的文学观念和评判标准。

文学研究论著目录的数据指标,可根据研究问题的需要而定。一般而言,文学研究论著需要下列指标:论著作者、篇名、文献来源(出版社或期刊名)、发表时间和被研究的作家、作品、文体及研究主题、研究方法等。前四项指标,属于客观的原生数据,即论著目录原有的信息;后五项指标,为主观的衍生数据,需要人工标引才能生成。因为有的论文和著作题目,没有明确或直接显示这些信息,需要人工识别后予以补充和标引。比如《“白发三千丈”新解》一文,如果不标引,计算机就无法识别这是研究哪位作家、哪篇作品、哪种文体,人工标引后,计算机就可以统计这是研究李白和他的诗歌《秋浦歌》。研究主题和研究方式,要根据自己的研究目的来自行设定,如果想统计分析田园山水诗、爱情诗等主题,那就要对属于这类主题的论著进行标引,不属于这类主题的论著就留空。对主题的体认,相对主观,同一论著,可以从不同的角度标引其研究主题和研究方式,最好是确定研究目的后再标引,否则容易造成数据指标的混乱。

作家基本数据的指标,包括姓名、生活年代、籍贯、作品、进士五大类。

姓名,应含字号、封号、谥号等信息。古人著述,习惯称作家的字、别号、封号、谥号而不称名。因此,数据指标中包含作家的字、别号、封号、谥号,既利于读者用户识别,也便于计算机从作品文本中挖掘提取相关数据。

生活年代,包括生年、卒年、享年数据,如有明确的享年信息就照录,没有享年记载,软件可根据作家的生卒年自动推算。生卒年信息不详的,可填写有明确纪年的时间点或大致的年代区间,以便确定其指数年。

籍贯数据,最好要有对应的古今地名。今地名至少分三级:省、市/州、县区,古地名需要与之对应的道或路、州/府、县三级。籍贯能细化到县级以下的乡镇就尽量细化。数据指标,越细分越好。有了省、市、县的分级数据,便于统计不同级别行政区域范围的作家分布。如果仅有市县级地名而无省级地名,或有省县级地名而无市级地名,需要统计分析相关区域数据时就会受到限制。文献记录的作家籍贯,往往只有一级地名或二级地名。古代行政区划地名的上下级关系及其与今地名的对应关系,计算机系统如果安装有相关的地理信息数据,可以自动比对处理,不需人工一一标注,但必要的核对工作是不可缺少的。

作品数据,包括别集名称、作品篇数。别集有不同名称的,尽可能将所有名称列入,以便挖掘数据和比对。作品篇数最好分体著录,如诗、词、曲、赋、文的篇数分别统计。记录篇数时还要注意题名和篇名的区别,因为一题多篇作品的现象十分常见。比如陆游,按诗题统计,他的诗歌只有六千多题,而按诗篇统计,则有九千多首。

进士出身,最好有进士及第的年份信息,未曾及第或及第时间不详的,只能阙如。由于学术研究不断进步深化,作家的生平事迹常常有新的考订成果予以订正补阙,因此,需要吸收最新研究成果来更新总集或别集里作家小传的信息。

作家活动和创作编年系地数据指标,可分活动时间、活动地点、活动内容和编年作品四大类。

活动时间数据,应包含帝王年号和干支纪年及对应的公元年份。如果数据库里有历代帝王年号、干支纪年和公元年份对照表,在数据处理时,可由系统自动生成对应的公元年或干支纪年,而不必人工录入。

活动地点数据,应包含古今四级地名,即古代的路、州/府、县政区和县级以下的景点、地名,与古代地名对应的今省、市/州、县和县以下的村镇或景点等,第四级地名有则录,无则留空。古代政区对应的今地名,最好由系统自动比对生成,这样不容易出现录入错误。当然,系统比对也会出现差错,需要人工校核和清洗。

活动内容信息,包括作家的活动经历和创作背景。活动经历需说明作家在其地是出生、寓居、仕宦

还是行旅、过访等因由;创作背景需说明因何而作、为谁而作。

编年作品数据,则需说明作品创作的起因、创作的时间(月日或季节)、创作的对象及交游者的姓名。如系与他人唱和,最好也列出他人唱和的作品篇名。文体属性也需标引,以便统计文体。

作品内容的数据指标,可分七大类:体裁、主题、人物、评点、研究、传播、接受数据。

体裁数据,即作品的文体分类数据,需要分层处理,第一层分诗、文、词、赋、曲等,第二层诗体下再分古、近体,第三层古体分乐府、歌行,近体再分律诗、绝句、排律等,律诗、绝句又分五言律诗和七言律诗、五言绝句和七言绝句等;词体既可按小令、长调来分,也可按音乐属性的令、引、近、慢来分。文章的分类最为复杂,《文选》《文苑英华》《唐文粹》《宋文鉴》《文章辨体汇选》《古文辞类纂》等文学总集分类各不相同。可考虑先按各种总集、别集原有的文体分类进行标引,然后进行汇总,看哪些作品没有文体分类信息、哪些作品的文体分类信息有差异,再进行人工统一处理。从文体分类信息数据,可以古人不同时代的文体观念。

主题数据,迄今无统一的划分标准。古代总集,有的是分类编排,如《文选》是先分文体,次分主题,文体下按主题分类,如赋体下分京都、畋猎、纪行、游览等主题,诗体下分劝励、公宴、祖饯、游览等主题;《文苑英华》赋体下按主题分天文、地类、岁时、京都、邑居、宫室、儒学、军旅、耕籍、人事诸类,诗体下分天部、地部、释门、道门、酬和、寄赠、送行等主题,每类主题下分若干子主题。这些主题分类,既是为古人创作时取资参考,也体现出他们的主题分类意识。这些分类数据,汇总之后可以分析古人的主题观念和文体体系、知识体系。每篇作品,先标引古人的主题分类数据,然后参酌今人的主题分类观念、知识体系重新划分。不同的分类原则和标准,划分出的主题类别和层次也大不相同。因而划分主题之前,可先确定主题分类标准体系,然后再细分相关主题或专题知识。比如诗文中写到的物象,可按现代知识分类,分动物、植物、矿物、器物、食物、衣物、景物等提取挖掘相关数据。

人物数据,指作品题目和本文中提及、描写的人物。人物有当世交往唱和的人物,也有提及评述的前代人物。这些人物数据,一部分可由计算机自动识别提取,一部分需要人工标引。姓名、字号、封号、谥号等人物同位语,计算机容易识别,但人物行第、官称等,计算机不容易识别。唐诗题目里常称人行第、职官,如杜二、高三十五、李侍郎、张吏部之类,计算机就无法确定是何人。这需要参考有关工具书来识别提取,如吴汝煜《唐五代人交往诗索引》,考实了唐诗题目中许多行第、官称所指人物。宋代诗文中的人物,还没有专门考释的工具书。古籍文献知识图谱平台(CNKG)已从历代诗文作品中提取了大量的人物数据,但部分官称如苏徐州、苏湖州、张吏部、王侍郎之类,平台还难以自动识别,只有通过别集笺注和作家年谱的相关考订成果来辨认。这类基础工作需要学者系统地梳理研究。人物数据,不仅要提取作家在作品中提及和评述别人的数据,也需要提取被别人的作品提及和评述的数据。双向挖掘人物数据,才能全面完整地呈现各种人物之间复杂多元的关系。

评点数据,包含两个方面:一是评人,二是评文。评人,是评论作家为人及其创作的数据;评文,指评论作家的具体作品。评文,涉及评次和评级两个维度。评次,指一篇作品被评点的次数。评级,指一篇作品被评价的等级。古人评诗衡文,常有优劣高低的定性评价。评次,好区分和统计,评级则需要人工来区分,因为古人对作品的评价,标准不一,用语说法不一。可考虑先将古人评点中的评价性用语全部汇辑,再区分其等级,并建立模型,看每个等级有哪些用语,然后再比对每篇作品所获品评的等级。

研究数据,指现当代研究古代文学作家作品的相关论文著作目录,而评点数据主要是指古代的评点资料和数据。研究论著数据,包含目录和论著文本两大类。作品的研究论著目录,前文已述及,主要是通过研究论著目录篇名来比对研究的是哪位作家的哪篇作品,如《李白〈蜀道难〉的主题新探》,计算机很容易识别这是研究李白的《蜀道难》,从而统计为李白《蜀道难》的研究数据。但许多论著本文中涉及的作品,并没有在论著题目中出现,如《李白评传》《李白研究》《中国文学史》等著作,题目上没有出现具体的作品篇名,但其文本内容却评论或谈及相关李白的作品,因此需要对现代研究论著的文本内容进

行挖掘分析,从中提出有关作品的评论数据。

传播数据,有五类:一为别集的版本数据;二为作品选的入选数据;三为非文学类典籍著录征引的数据,如史部、子部著作的著录与征引;四为书画传播数据;五为今日互联网、多媒体的传播数据。别集版本的多少能反映作家在历史上受欢迎的程度。宋代有千家注杜诗、五百家注韩柳文,都表明杜诗、韩柳文广受欢迎,故刊刻本甚多。一首作品被入选的频次越高,也表明其知名度、美誉度越高,一首没有选本入选的作品,很难有知名度和影响力。诗文作品,常常被文学之外的典籍著录和征引,如方志、笔记、佛书、道藏常有引用。这些都会扩大作品的影响。如杜甫《登高》诗"无边落木萧萧下,不尽长江滚滚来"①,宋代孙奕《示儿编》卷二、清人清斯《百愚斯禅师语录卷》第九和王士禛《古夫于亭杂录》卷二有引述评论。非文学典籍引录传播文学作品的数据,现在有了众多数字化古籍平台,查询和挖掘都比较方便。古籍文献知识图谱平台(CNKG)已将唐宋诗词作品与《四库全书》中的古籍做了关联,可查询和统计每首唐宋诗词被各类古籍引录传播的数据。一首作品被书法家书写成书法作品、画家画成绘画传播,也会扩大作品的传播力和影响力。作品被书画家书、画的频次越高,其知名度和影响力就越大。文物出版社 2001 年出版的段书安编《中国古代书画图目索引》、上海书画出版社 2005 年出版的卢辅圣主编《中国书画文献索引》都有作品篇名索引,可据以挖掘相关数据。

作品接受数据,也有多种,如诗词作品有后人追和次韵、引用和化用其中诗句的数据,有被后世小说戏曲引用改写的数据。追和、引用、化用的频次越高,表明其影响力越大、接受度越高。文章有仿效、改写的数据,如范仲淹《岳阳楼记》被南宋林正大隐括成《水调歌头》词,②还有引用的数据,比如,范仲淹《岳阳楼记》,影响后世文学创作自不待言,文章中倡导的先忧后乐的崇高精神更影响深广,仅在《全宋文》检索,提及、引用"先天下之忧而忧"的有 19 篇,在《全宋诗》和《全宋文》中用"先忧后乐"之句的有 61 次,在古籍文献知识图谱平台(CNKG)的古籍库里查询,用"先天下之忧而忧"的有 267 次,用"先忧后乐"句的有 313 次。这类数据,可考察一篇作品、一种观念、一种精神的影响力。

评点、研究、传播、接受的数据指标,又包含四个方面,一是评点研究、传播接受的对象即被评作家的姓名和被评作品的篇名,二是评点研究、传播接受的主体即评点者、研究者、传播者、接受者的姓名和书名或篇名,三是评点研究、传播接受的内容,四是评点研究、传播接受著作的版本及出版时间。确定了数据指标,就可以建模进行转化,然后建库进行自动统计。

三、数据意义

我们讨论了数据指标,明白了古代文学研究需要哪些数据、可以挖掘哪些数据,再来分析各类数据指标有什么意义、能发现和说明什么问题。

文学古籍书名目录数据,即古籍集部的书目数据,需要融合作家的基本信息数据进行分析。有了作家的生活年代和籍贯数据,就可以统计分析集部书目的时代分布与变化,看哪个时代、哪个时段集部的书目较多、哪个时段集部书目较少,其涨落变化有无规律可循、变化的原因又是什么;还可以统计并比较分析集部书目和经部、史部、子部书目的地域分布及其差异变化,可以比较南方和北方书目数量的地域差异及其变化。中国古代文学,一直存在南北差异,刘师培有《南北文学不同论》专文论述。③ 古籍书目数据是否反映出南方和北方作家有明显的差异?有待分析和验证。

集部书目,一般分总集和别集两大类,总集又分文选、历代、郡邑、氏族、唱酬、题咏诸属。二、三级类属数据,也有学术意义。如郡邑类的总集数据,可以结合作者的籍贯数据,分析相关地区的文学风气、文

① 仇兆鳌注:《杜诗详注》卷二十,北京:中华书局 2010 年版,第 1766 页。

② 林正大:《水调歌头》,唐圭璋:《全宋词》,北京:中华书局 1965 年版,第 2451 页。

③ 参程千帆:《文论十笺》,哈尔滨:黑龙江人民出版社 1983 年版,第 81—125 页。

坛好尚;氏族类总集数据,可以分析家族的文学氛围和文学成就;唱酬类总集数据,可以分析考察诗词唱和之风的形成与演变等。

作品篇名目录数据,可以从不同的范围、不同的维度来分析。断代总集的作品目录数据,可以分析文体的发展变化。如用《全唐文》和《全宋文》的篇目数据分析唐宋两代文体的分布及变化,看唐宋时代哪些文体的作品较多、哪些文体较受重视,是应用性的公文多还是审美性的文章多,进而分析有关文体的应用场景;还可以考察唐宋时代文体的演进,宋人新增了哪些文体,哪些文体式微甚至消失。这些都是传统的定性描述很难发现的问题。至于《全唐诗》《全宋诗》的作品目录,可以用来分析不同阶段诗体的冷热变化,不同时期、不同个体在诗体的选择运用上有什么差异和变化,诗体的不同选择是否体现出诗人不同的创作个性和审美好尚。诗体数据,又可以分析验证平仄规则、体式特点,比如,律诗的三平尾和三仄尾是否合乎规则要求,历来各执己见。我们统计了唐宋两代创作五七言律诗 30 首以上的诗人 534 位、诗作 104174 首,出现过三平尾的律诗只有 177 首,占统计范围内律诗总量的 0. 17%;而含三仄尾诗句的律诗则有 8143 首,占统计总量的 7. 81%。三仄尾出现的频次比三平尾要高出四十多倍。这表明三仄尾的限定相对宽松,而三平尾的限定相对严格。统计数据可以验证相关格律规则的普适性和宽严度。《全唐五代词》《全宋词》的调名数据,则可用来统计各词调的使用频次、词调的生长变化,分析不同风格流派的词人用调差异等。这方面,刘尊明教授做了不少的探索,取得了一系列成果①。

诗文选本的目录数据,可以用来分析作家作品的影响力。一个作家入选的作品越多,其影响力越大;一篇作品被不同选本入选的次数越多,其影响力就越大。我们曾以唐诗选本和宋词选本的入选率等数据指标,量化分析唐诗和宋词经典名篇的影响力,发现了不少有学术意义的问题,也提出了一些传统定性分析无法得出的结论②。选本选谁不选谁、选哪位作家多少作品、选哪些作品,都反映出编选者及其所处时代的文学观念、审美标准。曹道衡先生曾指出:“从来的编选者都不可能任意决定他的文学观和选录标准。因为任何人的文学思想都不是凭空产生的,而是在吸取和改造前人学说的基础上形成的。具体到总集的编选来说,编选者对前人众多的作品究竟选录哪些、不选录哪些,这虽然由他的文学观来决定,但他的文学观本身却无法完全摆脱传统的影响。早在他文学观形成之前,他已从接受传统的过程中形成了某一类作品是好的、某一类是较差的;某些作品是名篇、某些则不是的想法。”③所以,借助选本的选目数据,可以量化分析不同时代不同选本的文学观念和选录标准及其变化,还可以通过标引选目的主题和题材,获得所选作品的主题和题材数据,分析不同时期的选本对不同题材和不同主题的偏爱和弃取。比如,20 世纪 50 年代以来,受当时文学观念的影响,许多反映社会现实和爱国主题的作品多被选本挖掘选入,如杜甫的“三吏三别”、白居易的《卖炭翁》之类作品入选率特别高,而在古代相关选本中则少见选录。总之,选本的作品目录也是非常有用的数据。

作品选目录数据能反映作家的影响力,文学史的章节目录数据则能显示作家地位的高低。什么样的作家独占一章,什么样的作家合为一章,哪些作家独占一节、哪些作家合为一节,哪些作家在章节目录里出现,哪些作家的名字不上章节目录而只在正文叙述中附及,都体现出文学史编写者的文学观念和价值标准,也反映出对作家地位和贡献的不同评价。正如钟嵘《诗品》将诗人分为上中下三品,文学史的章节目录也同样体现一种品第观念。独占一章的作家都是一流的伟大作家,如屈原、李白、杜甫、苏轼、关汉卿等,独占一节或合为一节的作家,都是在文学史上有重要影响的著名作家。同一作家的地位,在不同文学史里是不同的,比如,晚唐诗人李商隐,在 20 世纪的大多数文学史里,只占一节的篇幅,而到高

① 刘尊明、王兆鹏:《唐宋词的定量分析》,北京:北京大学出版社 2012 年版。

② 王兆鹏、孙凯云:《寻找经典——唐诗百首名篇的定量分析》,《文学遗产》2008 年第 2 期,第 40—52 页;王兆鹏、郁玉英《宋词经典名篇的定量考察》,《文学评论》2008 年第 6 期,第 79—86 页。

③ 曹道衡:《汉魏六朝文精选·前言》,北京:商务印书馆 2018 年版,第 2 页。

等教育出版社出版的《中国文学史》①,李商隐就赫然独占一章,表明李商隐的诗史贡献被提高到与李白、杜甫同等位置。如果统计 20 世纪以来各种中国文学史的目录数据,就可以有效分析和观察古代作家地位的升降变化。而作家地位的变化,又反映出文学史撰写者及其时代的文学观念变化。

论著目录数据,可分别从时间、地域、作家、文体、期刊、论著作者等角度来统计分析。时间角度,可统计分析一个研究领域不同时段的选题分布、文体选择、作者队伍的不同特点。笔者就量化分析过 20 世纪国内唐代文学研究的发展历程和词学研究的阶段性变化。② 地域角度,可探讨国别或地区有关某个断代文学或某种文体或某个作家的研究状况及其变异。我们曾探讨 20 世纪日本和韩国有关唐代文学研究成果量的发展变化。③ 作家角度,可统计分析一位作家研究的选题角度、研究面向和研究历程等,如 20 世纪有关李白和杜甫的研究,已有论文进行统计分析。④ 文体角度,可探寻特定时段有关某种文体研究的选题分布与发展历程,如唐代小说研究已有相关成果做过考察。⑤ 期刊角度,可研究一种期刊的选题分布与作者队伍的构成情况。我们团队曾就《文学遗产》《阿来研究》两种期刊的选题分布与作者队伍做过统计分析。⑥ 作者角度,可统计分析特定时期相关领域研究论著作者的规模、成果量级、年龄梯队等。论著目录数据分析,属于计量学术史研究,其理论方法还有待完善,发展空间十分广阔。

作家基本信息数据的每种指标,各有各的作用和意义。

作家的生活年代和作品数据,可以统计分析作家的时代分布,看每个时期作家和作品数量有什么变化、文学的高峰期和低谷期是在哪个时段、作家和作品数量的变化跟哪些因素有关,有无规律可循。

作家的年寿数据,可以考察不同时代作家的平均寿命和创作年限。有学者根据宋人文集和出土碑刻墓志的 1765 个样本统计,宋代士人的平均寿命为 61.68 岁⑦。笔者根据《全宋诗》中有生卒年可考的 2423 位诗人的年寿数据统计,宋代诗人的平均寿命为 65.03 岁,高于一般士人的平均寿命。另有学者根据唐代 5053 方墓志统计,得出唐人平均寿命为 59.25 岁的结论。⑧ 笔者根据唐代有年寿可考的 383 位诗人统计,唐代诗人的平均寿命为 64.35 岁,也高于一般唐代人物的寿命。当然,唐代诗人年寿可考的数据有限,会影响到数据的代表性。但至少提供了不同的数据样本,据此可以探讨为什么诗人的平均寿命比一般人要高。

作家的籍贯数据,则可考察作家的地域分布及其在不同时段的变化,看各个地区在什么时代文学比较发达、产生的作家作品较多,考察地区文学发展是均衡性大于失衡性还是相反。比如福建,在唐代作家人数还比较少,到了宋代作家人数激增。我们根据《全宋词》《全宋诗》《全宋文》开发的作者数据库统计,北宋时期,按路(省)分统计,福建路的作家人数,名列第二,共 1682 人,仅次于第一名的两浙路;如按州府统计,位居前两名的则是福建路的福州和建宁府:福州有 542 位作家,高居首位,建宁府(建州)有 378 人,位居次席。这很出乎我们的意料。因为在我们的印象中,宋代江西、江苏、浙江的文学很

---

① 袁行霈、罗宗强主编:《中国文学史》第二卷,北京:高等教育出版社出版 1999 年版。

② 王兆鹏:《20 世纪国内唐代文学研究历程的量化分析》,《文学评论》2015 年第 4 期,第 184—197 页;王兆鹏:《20 世纪词学研究成果量的阶段性变化及其原因》,《学术研究》2010 年第 6 期,第 127—135 页。

③ 王兆鹏:《20 世纪日本唐代文学研究成果量的发展变化》,《社会科学战线》2015 年第 5 期,第 142—150 页;王兆鹏:《20 世纪下半叶韩国唐代文学研究成果的量化考察》,《贵州社会科学》2014 年第 8 期,第 42—46 页。

④ 王兆鹏、戴峰:《20 世纪海内外杜甫研究成果量的时段变化》,《江海学刊》2015 年第 6 期,第 186—193 页;郭红欣、谭新红《20 世纪国内李白研究成果量的量化分析》,《江海学刊》2015 年第 6 期,第 194—199 页。

⑤ 王兆鹏:《20 世纪中外唐代小说研究成果的的量化分析》,(台北)《文与哲》第 27 期(2015 年 12 月)。

⑥ 王兆鹏:《〈文学遗产〉复刊以来论文和作者队伍的统计分析》,《文学遗产纪念文集》,文化艺术出版社 1998 年版;王兆鹏、赵黔:《〈阿来研究〉十年来的选题分布与作者队伍的统计分析》,《阿来研究》第 20 辑(2024 年)。

⑦ 吴志浩:《宋代士人平均死亡年龄考》,《浙江学刊》2017 年第 4 期,第 170—181 页。

⑧ 蒋爱花:《唐人寿命水平及死亡原因初探——以墓志资料为中心》,《中国史研究》2006 年第 4 期,第 59—76 页。

发达,没想到福建的文学也会如此昌盛。

进士数据,也有助于统计分析哪些地区进士人数较多、文化教育发达。我们根据龚延明、祖慧《宋代登科记考》转化的数据统计,宋代州府这一级进士人数最多的也是福州,达 4681 人,其次是兴化军(今福建莆田)1769 人、泉州 1473 人,前三名被福建的州府包揽。福建建州(今福建建瓯)965 人,位列第八。进士人数多,反映出当地教育水平的发达。而文学创作的繁荣度是与教育的发达度成正比的。宋代福建的福州、建州作家人数众多,与其教育水平高度正相关。① 作家小传的信息数据,虽然比较单纯而不繁杂,但功用却不小,可以发现不少文学史以前没有关注到的现象和问题。

作家活动和创作编年系地数据,也可进行不同角度的分析。

从宏观角度,可统计分析一个时期文学发展变化的历程。比如,北宋元祐时期是宋代文学的高峰。近人陈衍就认定唐宋诗的三个高峰时段分别是盛唐开元、中唐元和、北宋元祐。数据显示,这三个阶段,无论是作品量还是大诗人的数量,都处在高峰状态。而元祐时期的作家活动和创作编年数据,不仅显示元祐文学是北宋文学创作的最高峰,而且还动态显示了元祐文学的发展进程。元祐元年、二年(1086—1087),高开高走,呈爆发式增长。与上一年元丰八年(1085)483 篇的作品量相比,元祐元年与元祐二年快速增长至 981 篇和 1162 篇;元祐三年至七年(1088—1092)的作品量仍维持在高位,元祐八、九年(1093—1094)作品量明显下滑,仅有 369 篇与 381 篇,约为高峰期元祐二年的三分之一,作品数量上呈现出明显的衰退。② 从地域分布来看,元祐时期的文学中心在京城,元祐九年,随着苏门文人集团和旧党人士被贬出京,文学的中心也移到南方的贬谪之地。

从微观角度,可统计分析个体作家创作阶段和空间分布。比如,苏轼一生创作十分丰富,哪个年龄段是他的创作高峰、哪个地方是他的创作高峰,无法进行主观的感悟式描述或例举式证明,只有统计数据才有说服力和客观性。有了苏轼一生活动和创作编年系地数据,就可以对苏轼创作的时空分布及变化进行量化分析。据统计,苏轼 40—59 岁的中年创作量最旺盛,年均创作诗词文 262 篇,这 20 年的作品量占了他一生作品量的六成半,而 40 岁以前的作品量约占一成半。从人生顺逆期来考察,苏轼贬谪时期的创作量是平顺期创作量的两倍多,黄州和惠州是其诗词创作的两座高峰。③ 谪居黄州和惠州时,虽然他一再声称畏祸不敢写作,好友也劝他不要再写,但创作的冲动无法抑制,实际上照常写作,只是不敢公开传播而已。他在惠州对友人说:“公劝仆不作诗,又却索近作。闲中习气不除,时有一二,然未尝传出也。”④创作,是苏轼生命的存在方式,而不仅仅是表现方式。

作家活动和创作编年系地数据,转化为文学编年地图后,还能发现诸多有意义的问题,比如一个作家的行程路线,往往反映出当时的公共交通路线。苏轼出川到开封,曾走两条路线。第一次出蜀走西线,经广元过剑阁入陕西,道洛阳到开封;第二次出蜀走中线,沿长江东下,在湖北荆州上岸陆行,经襄阳、洛阳到开封。其后乃父苏洵在汴京去世,苏轼兄弟俩扶柩回乡,走的则是东线,先由汴河南下至楚州,再沿运河到瓜洲渡,由长江西上入川。西线全程走陆路,是唐宋时期入蜀的主要干道,唐玄宗入蜀走的就是此条路线。中线是一半水路一半陆路,杜甫《闻官军收河南河北》“即从巴峡穿巫峡,便向襄阳向洛阳”⑤,规划的出蜀回乡路线,也是这条道路,只是他后来未能成行。东线全程为水路,欧阳修贬谪夷陵(今湖北宜昌),走的路线跟苏轼兄弟扶柩回乡路线前段完全相同,只是到夷陵而止,没有进三峡入蜀;范成大和陆游由江

---

① 王兆鹏、齐晓玉:《宋代诗词文作者队伍的层级与时空分布》,《中南民族大学学报》2022 年第 1 期。按,此文的作者地域分布数据,是按当今行政区划统计,故与本文的统计数据有差异。

② 高武斌、邵大为:《北京元祐文学的图景特征——基于〈唐宋文学编年地图平台〉编年系地作品的数据统计》,《中南民族大学学报》2023 年第 5 期,第 146—155 页。

③ 郭红欣:《苏轼作品量的时空分布》,《中南民族大学学报》2020 年第 1 期,第 112—118 页。

④ 孔凡礼点校:《苏轼文集》卷五十八《与曹子方五首》其三,北京:中华书局 1986 年版,第 1775 页。

⑤ 仇兆鳌注:《杜诗详注》卷十一,第 968 页。

浙入蜀,也是先由运河再沿长江西上,后段入蜀路线也与苏轼回乡路线相同。

作品中的人物数据,可分析一个作家跟前代各种人物的关系,也可分析其与同代作家的关系,还可分析其与后代作家、批评家的关系。古籍文献知识图谱(CNKG)的人物栏目可自动呈现唐宋时代每位作家与前代、同代、后代各种人物的关系,并能自动呈现每个作家可视化的社会关系网络图。据此,既可以统计分析一个作家所受前代历史人物的影响,也可量化考察他与同代作家的关系及其对后世的影响,还可分析各种历史人物的文学影响力。这方面的大量数据亟待利用。

作品内容的其他数据指标,因暂无完善的数据可资利用,其意义容后申论。

文学研究数据,可以发现新问题、解决新问题,但并不意味着什么问题都可以解决。统计数据和定性资料一样,也有它的局限性。定性分析与定量分析的有机结合,才能最大限度地解决文学研究的各种问题。

Data Sources, Indicators and Their Significance in the Research on Classical Chinese Literature

Wang Zhaopeng

(*School of Literature, Journalism and Communication, South-Central Minzu University, Wuhan* 430074, *China*)

Abstract: The research on classical Chinese literature requires datafication, with bibliographic catalogs and texts of works serving as data sources. Catalogs of book titles, titles of works, chapters and sections in literary history, and works of literary research can all be transformed into literary data. Elements such as literary genres, themes, characters, commentaries, research, dissemination, and acception within the texts of works can also be extracted as data. As diverse types of data carry different significances, the directory data from collected works of a particular dynasty can be used to analyze the evolution of literary styles. The directory data from prose and verse anthologies can be used to assess the influence of an author's works. The directory data of chapters and sections from literary history can reflect the status of writers, as well as the literary views and value standards of compilers, which may change over time. The directory data from academic writings can be statistically analyzed to track the development process of academic history. The data regarding authors' lifetime and their works can reveal the spatiotemporal distribution of authors, while the data concerning the characters within the works can help explore authors´ social relationships and their influence. However, the data of literary research may have their limitations. An organic combination of qualitative and quantitative analyses is essential to the effective solution to various issues in literary research to the maximum extent.

Key words: research on classical Chinese literature; digital humanity; quantitative analysis; datafication

(责任编辑　郑　园)